2018

A Quantitative analysis of the Kenyan students' loan default

Pauline Nyathira Kamau
Strathmore Institute of Mathematical Sciences (SIMs)
Strathmore University

A Quantitative Analysis of the Kenyan Students' Loan Default

By

Pauline Nyathira Kamau - 093353

This research project is submitted to the Strathmore Institute of Mathematical Sciences in partial fulfillment of the requirement for the degree of Masters of Science in Mathematical Finance.

STRATHMORE UNIVERSITY

April 2, 2018

Supervisor: Dr. Lucy Muthoni
Co-supervisor: Dr. Collins Odhiambo

*DECLARATION

I, the undersigned, declare that this study is my original work unless otherwise stated and to the best of my knowledge has not been presented for credit in any other University.

Sign..... **Date**.....

Kamau Pauline Nyathira
Reg. No. 093353

This project has been submitted for examination with my approval as the University Supervisor.

Sign..... **Date**.....

Dr. Lucy Muthoni
Institute of Mathematical Sciences, Finance
Strathmore University

Sign..... **Date**.....

Dr. Collins Odhiambo
Institute of Mathematical Sciences, Statistics
Strathmore University

*DEDICATION

This project is dedicated to my mum (Magdaline Wambui), dad (Gabriel Kamau) and my brothers for their undying love and support. I also extend gratitude to my friends for their moral support, and above all the Almighty God for His provisions that were sufficient throughout my studies.

*ACKNOWLEDGMENT

I would like to thank each and every one who assisted my participation in the masters program at Strathmore University and in compiling this project. While it is not possible to thank everyone by name, I would like to extend special thanks to Dr. Lucy Muthoni and Dr. Collins Odhiambo (data analysis and Statistics supervisor) for their invaluable guidance, support and assistance through information provided regarding pertinent issues related to the program and this project.

I also wish to thank HELB recovery department fraternity for providing me with the data, and allowing me to use their organization as a case study.
Special thanks goes out to my family for their patience, motivation, guidance and support. I would also like to extend gratitude to my classmates for the shared experiences and valuable contributions, as well as the friendships that were forged among us. I hope the friendships will endure the test of time as we grow professionally.

Above all, I am grateful to the Almighty God for good health, well being and sound mind that enabled me to see this program through.

List of Figures

*LIST OF ABBREVIATIONS

1. HELB - Higher Education Loans Board
2. HELF - Higher Education Loans Fund
3. USA - United States of America
4. USSR - Union of Soviet Socialist Republic
5. KRA - Kenya Revenue Authority
6. PIN - Personal Identification Number
7. JAB - Joint Admissions Board
8. KUCCPS - Kenya Universities and Colleges Central Placement Service
9. VBA - Visual Basic for Applications
10. AIC - Akaike Information Criterion

Abstract

Higher education capacity, quality, and availability has driven more countries to turn to student loan schemes in order to assist students whose families are unable to meet their university costs. Ideally, all students seeking university education should be able to access these loans. It is also expected that student loan applicants pay back the entire loan in the stipulated time frame to allow other needy students joining university to utilize the repaid amounts.

In this study, we seek to perform a quantitative analysis of loan applications by computing the probability of default of a given applicant using the qualitative information provided in the application forms. We apply multiple logistic regression with the binomial nominal variable defined either as defaulter or re-payer. Further, we treated different factors affecting default probability of the student as independent variables. The main objective was to find out the effect that the independent variables have on the dependent variable. We then validated the resulting model by comparing its results to observed data from the Kenyan Higher Education Loans Board.

Results show the amount of loan reimbursed as the main factor affecting default. This can be an eye-opener for policy makers in their effort to mitigate non-repayment.

Keywords: **Student loans, Default rates, Multiple Logistic Regression**

Contents

1 Introduction

A student loan is designed to assist students pay for university or college education and expenses that are associated with it, such as tuition fees, purchase of books and stationery, hostel/rent expenses among other living costs. It usually differs from other types of loans in that the interest rates are much lower, and payments are deferred until one year after students' successful graduation.

Loan default refers to the failure to meet the obligation or condition of paying back the loan. It is very important to understand student loan default patterns and the factors that lead to default because it has caused a huge economic crisis for graduates as well as new university entrants. Student loan default is usually accompanied by other competing events such as, the question of whether it is the first time that the individual has borrowed and defaulted, or if the individual borrowed several times and defaulted once or they have defaulted on all occasions. An important factor is to find out whether default is a single occurrence or is a recurrent issue given a sample of students' information. Analysis is usually carried out using Cox regression model to investigate the time it takes to occurrence of an event of interest, such as default.

This study focuses on the first time the student defaulted given several variables. We determine factors affecting loan default. Student loan default affects all of the stakeholders involved, among them the students, Kenyan Higher Education Loans Board (HELB), the government through the ministry of Education, Science and Technology and the economy at large.

Data provided by HELB is qualitative in nature and is provided by loan applicants at the point of application. It contains information about the student's background and parent's employment status among other details. We shall only consider information that is relevant to our study which we shall use to explain how the different variables lead to default.

Numerous studies have been done concerning student loan default using different models and methodologies,(Herr and Burt, 2005; Jacob P.K. Gross et al., 2010; Woo J.H., 2002).

This study explains this matter specifically by use of Multiple Logistic Regression which will have an outcome that will tell us if the individual either defaulted (1) or did not default (0) on their loans. We then confirmed that our model is correctly specified and relevant by use of several tests to ensure unbiasness, consistency, test the variance inflation properties among other tests. Then, we interpreted the results and discussed what they meant for Kenyan student loan applicants and for the Board especially concerning its loan disbursement policies.

1.1 Background of the study

In Kenya, students who are not able to fund their education are enabled to do so by existence of HELB. It is the main source of university and college financing in Kenya. It was established by an Act of Parliament (Cap 213 A) in 1995 under the then Ministry of Higher Education, Science and Technology. The main purpose of the Board is to disburse loans, bursaries and scholarships to students pursuing higher education in recognized institutions. The Board's roots date back to 1952 when the colonial government awarded loans under the then Higher Education Loans Fund (HELF) to Kenyans pursuing higher education in universities outside East Africa notably Britain, USA, former USSR, India and South Africa. These students were not on scholarships and advanced loans by the government at that time against land title deeds and insurance contracts.

HELB was meant to establish a Revolving Fund from which funds could be drawn to lend out to needy Kenyan students pursuing higher education. The establishment of a revolving fund was also expected to ease pressure on the exchequer in financing education, which currently stands at 40 per cent of the annual national budget (Source: HELB database, Loan Repayment and Recovery).

1.2 Problem Statement

The implications of student loan default are in abundance. For instance, the treasury incurs huge losses from non-repayment of the funds it provides to HELB. This is because the Board does not recover the loans as efficiently as it should. In such a case, the

greatest losers are future university entrants who will likely miss out on the opportunity to join university due to failure of the regeneration of the revolving fund caused by defaulters. The rising number of borrowers who cannot pay back their education loans have raised questions on the efficiency of disbursement and recovery of student loans in Kenya.

Institutions of higher education also incur losses when new entrants are unable to secure the loans and thus are unable to pay their tuition fees. This creates a group of individuals who have no means of getting their education and have to enter into informal employment which as well creates a whole other cohort of problems for the economy, all due to loan default. Student loan default might be associated with numerous systematic patterns, which if pinpointed, may help design public policy interventions and strategies, thus reducing the likelihood of defaulting (Muthii, 2015).

1.3 Main objectives

- To determine factors which play a key role in increasing default rates.
- To develop a quantitative model that returns an individual's risk of default (A risk-profiling model).
- Suggest ways in which default rate can be decreased.

1.4 Significance of the Study

Higher education is important because knowledge and skills imparted at this level are crucial for the transmission of core values, which exposes one to more opportunities for both self and national economic growth. In Kenya, surveys show that only 60 per cent of students are able to access government student loans. At the same time, KSh. 20.04 billion has not been paid up by those who were granted the loans. Of concern, non-performing loans currently make up more than 50 percent of the outstanding loans (Source: Kenyan Higher Education Loans Board). The poor performance of beneficiaries has led to a huge decline in the revolving fund which is used to finance new university entrants.

Several studies suggest that students who go to two-year course institutions portray higher default rates than their peers at four-year public or private institutions (Podgursky, Ehlert, Monroe, Watson, and Wittstruck, 2002; Woo, 2002), even when the time horizon for considering default is extended to six years (Kesterman, 2005). Moreover, greater institutional investment and instructional support is associated with decreased likelihood to default (Volkwein and Szelest, 1995). Generally, the wealthier the institution attended and the greater the student's access to social and economic capital, the less likely the student is to default (Hillman and Hossler, 2009). This clearly suggests that an institution's characteristics play a role on whether a student may or may not default.

It is important to note that adequate financing of higher education can only be achieved if meaningful collaborations are established between all the major stake-holders including the government, through the Ministry of Education, Science and Technology, parents and guardians, students, employers and institutions of higher learning as well as other financiers of higher education (HELB, Annual Report and Financial Statement, 2015). The main source of the Board's funding is meant to come from repaid loans from former students, and even though realistically speaking the Board cannot recover 100 percent of loans disbursed especially due to unemployment of graduates, it should be able to recover more than 85 percent of the loans. This is, however, not the case. All the above-mentioned stakeholders are affected in more than one way when default occurs. An additional affected stakeholder is the general performance of the economy where the cost of education does not correspond to the benefits expected from funding it.

HELB should take stricter measures when risk profiling potential beneficiaries in order to make loan recovery strategies easier, and at the same time minimize the current default rates. One of the suggestions floated so far is that students should apply for their Kenya Revenue Authority (KRA) personal identity number (PIN) immediately they turn the age of eighteen, as they do for national identity cards. This will allow the KRA to track tax-related financial transactions that the individuals engage in, providing background information at the point of application for HELB loans. This would serve well as a reference for the student's credit score.

Another suggestion is that HELB should seek out childhood behavioral information concerning the loan applicants. Any information about child delinquents should be recorded by primary and secondary school teachers, and this information used to determine whether such behavior (if damning) increases the probability of loan default by beneficiaries. This will reduce default and ensure that loans are being disbursed to individuals with higher probability of paying back the loans.
This study fronts a more quantitative method of reducing default rate by looking at different factors that affect an individual's ability to repay the loan. Reducing default rate means that more funds will be available for those students who cannot afford to finance their education.

2 Literature Review

This chapter reviews the literature into the factors which are believed to contribute to default of higher education student loans, which for the most part are personal details about the applicants or their background information (Hillman, 2014). The role of personal characteristics in student loan default are examined as well as socio-economic factors, education experience and post university experience.

2.1 Personal Characteristics

Age Studies find that age has a positive relationship with default i.e. as one grows older, their probability of default increases. According to Woo(2000), as one gets older then one is more likely to default than when he or she is younger, perhaps due to a weakening of dependency to parents and family who might assist the student while experiencing financial difficulties. Default by older beneficiaries can also be explained by increased financial obligations that come with growing older, for example, career obligations and family support which may slow down loan repayment (Herr, 2004).

Additionally, older beneficiaries may be more likely to default because they owe more than their younger counterparts as a result of interests accumulating over time and may

have relatively less in available resources to repay the loans. Interests for HELB accumulate by KES 5,000 every month if the loan is not continuously serviced or if repayment stops at a given time during loan servicing. This may become a huge burden as one gets more older.

Gender There exists a positive relationship between gender and loan repayment. Female borrowers are inherently slightly less likely to default compared to male borrowers. These findings agree with a number of recent studies that have shown that men are more likely than women to default on student loans (Woo, 2002). Women appear to shy away from debt compared to men, according to an Experian analysis released in May, 2013. The study found that, on average, men carried 4.3 percent more debt than women. On a general note, their mortgages were 4.9 percent higher than home loans taken out by women. At the same time, women tend to use less available credit on their credit cards than men — 30 percent versus 31 percent.

While men may seem to be more comfortable taking on more debt, they also get into financial trouble more often. The Experian analysis released in May 2013 found that men were more likely to fall 60 days or more behind on their mortgage payments than women — 5.7 percent versus 5.3 percent. *"Typically, a mortgage is the largest debt that an individual has in a lifetime,"* says Rod Griffin, a director of public education *"the way you manage that is very telling of your financial situation."* (www.bankrate.com, 22/01/2018, Janna Herron).

Marital Status Family structure can affect the likelihood of defaulting on loans in a number of ways. Being single, divorced or widowed was found to increase the probability of defaulting by more than 7 percent (Volkwein and Szelest, 1995). Being a single parent was also associated with a greater risk of loan default (Volkwein et al., 1998).

2.2 Social-Economic Factors

Dependents The greater the number of dependents claimed by a student, the greater the likelihood of loan default (Dynarski, 1994; Volkwein and Szelest, 1995; Woo, 2002).

Volkwein and Szelest (1995) found that the probability of default increased 4.5 percent per dependent. This can include children or younger siblings. As common sense suggests and research has collaborated, having more dependents requires a greater share of one's finite supply of resources, thereby decreasing the ability of a student with dependents to repay loans (Herr and Burt, 2005). Indeed, this was found in one study to have a greater effect on the likelihood of loan default than the type of institution attended, parent's income, and even the student's annual earnings (Volkwein et al., 1998).

Parents' Education Level A large body of research has found that, given the positive relationship between education and socioeconomic status, students whose parents had higher levels of formal education were less likely to default than first generation college students (Choy and Li, 2006; Volkwein et al, 1998; Volkwein and Szelest, 1995). This is true in relation to the mother's as well as the father's level of education (Steiner and Teszler, 2005).

Parental Income/ Employment status As we would expect, students from low-income families tend to incur more debt during school than their wealthier peers (Herr and Burt, 2005; Steiner and Teszler, 2005; Volkwein and Szelest, 1995). Low-income students also report feeling more burdened once their loan repayments begin, and some evidence suggests this reaction is intensifying (Baum and O'Malley, 2003b).

Generally, the higher the family income the lower the likelihood that the student will default (Knapp and Seaks, 1992; Wilms et al., 1987; Woo, 2002). Families with more money are able to provide a financial safety net unavailable to students from lower-income families, who are more likely to need such a resource given their greater levels of riskiness. This safety net also helps students to meet their loan obligations through fluctuations in personal income.

2.3 Education Experience

Degree or Major chosen Studies show that the student choice of a degree major has a positive relationship with loan repayment. Harrast(2004) states that some majors tend

to be more resilient to the labor market conditions than others. For example studying special education, computer engineering, sociology, art history, or risk management and insurance is associated with higher levels of debt relative to other fields. His study focused on one institution, however, and the author was unsure why a student's choice of major affected subsequent debt burden, Hossler and Hillman (2010).

More evidence exists to suggest that post graduation earnings related to field of study affect personal income and therefore, one's ability to repay loans (Herr and Burt, 2005; Steiner and Teszler, 2005). Lochner and Monge Naranjo (2004) found the effects of major choice disappeared after controlling for total debt and post college earnings. According to Flint(1997), the greater the incompatibility between a student's undergraduate major and his or her field of employment, the higher the risk factor for default.

Bursaries/ Scholarships Awarded The Kenya Universities and Colleges Central Placement Service (KUCCPS) is a corporate body established under the Universities Act 2012 to succeed the Joint Admissions Board (JAB). The service is governed by the placement board which also seeks to establish a criteria to enable students access the courses for which they applied taking into account the students' qualifications and listed priorities (Source: KUCCPS.net).

The relationship between sponsorship application and default of student loans has not been studied extensively in Kenya as per the available literature.The study done so far compared bursary and loan applicants' default rate against those that applied for loans alone. The results showed that bursary applications have a negative relationship with loan repayment implying that bursary applications are at best suited to predict probability of loan default.

2.4 Post-University Experience

Employment Status The ease of finding formal employment in Kenya is at an all time low in the year 2018. Many graduates are finding their way into informal employment which is making it even harder for HELB to track these individuals for loan repayment.

Not to mention that increase in informal employment has led to a substantial decline in the overall economic growth in Kenya because the people are living hand to mouth with very futile plans for future investments.

If one is in informal employment, a brief illness or injury usually leads to the employee getting dismissed and replaced by other people. There is therefore a lack of job security for individuals and this makes it impossible to focus on paying back student loans. These individuals tend to as well be below the taxable income bracket, thus adding onto the fact that they are unable to repay their loans, the government cannot get anything from them tax-wise in order to fund HELB.

In this study, we will look at the probability of default given most of the factors studied as affecting default rates being present.

3 Methodology

3.1 Introduction

The main purpose of this paper is to identify the major factors that explain what causes student loan default. The analytic technique of choice is Multiple logistic regression given its ability to predict a nominal dependent variable from one or more independent variables. Logistic regression is one of the statistical models in generalized linear models.

Logistic regression lets us predict a discrete outcome, such as membership or involvement in a group, from variables that may be continuous, discrete, dichotomous, or a mix of any of these. The dependent or response variable in this case is dichotomous, meaning it can only give us one of two outcomes, whether a student is a defaulter or a re-payer.

We have several independent (X) variables: gender, age, degree or major chosen, marital status, number of dependents, bursaries or scholarships awarded among other variables. These variables are believed to play a role in the determination of the dependent (Y) variable. The purpose of a multiple logistic regression is to find an equation that best

predicts the Y variable as a linear function of the X variables and elaborate to what extent these independent variables affect the dependent variable.

We tested the null hypothesis for the X variables, that adding them into the multiple regression does not improve the fit of the equation any more than it would by chance. We used the p-values attained for the null hypothesis as a guide to build the multiple regression equation. A study by Schwab, (2002) showed that sample size guidelines for multiple logistic regression indicate a minimum of 10 cases per independent variable and continues to minimize the effects of outliers and influential cases as the sample enlarges.

Multiple logistic regression is often considered an attractive method of analysis because it does not assume normality, linearity, or homoscedasticity. Another alternative to multiple logistic regression is discriminant function analysis which requires these assumptions to be met, (Starkweather and Kay, 2012).

Discriminant analysis can as well be used to predict membership or involvement with only two outcomes. However, it can only be used with continuous independent variables. Thus, in an instance like this one where the independent variables are a mix of continuous and categorical, logistic regression was the most preferred for analysis.

In spite of the differences in assumptions, Pohar et al., (2006) and Antonogeorgos et al., (2009) argued that, the differences between logistic regression and discriminant analysis methods become negligible if the sample size was enlarged to a certain value, say 100 observations or more. Additionally, a study by Cleary and Angel (1984) showed that both methods often have similar results.

3.1.1 Exploratory Analysis

This is a general overview of what the data looks like. An exploratory analysis is important because it will help us understand the kind of information in the data before we get into the model. The entire sample includes 5,100 individuals.

The table below shows HELB's interest rates and the frequency of students in our sample who pay the rates. Those who graduated between 1974/75 and 1994/95 academic years repay their loans at 2 percent, while those who took loans from 1995/96 to date repay their loans at 4 percent. HELB can vary the interest rate anytime without referring to the beneficiary (Section 6(c) of the HELB Act). For postgraduate and continuing education students, the interest rate is 12 percent compounded annually, (http://www.helb.co.ke/loan-repayment/, 26/04/2018).

Interest rate	Frequency	Percent	Valid percent	Cumulative percent
2	2	.0	.0	.0
4	4821	94.5	94.5	94.6
12	277	5.4	5.4	100.0
Total	5100	100.0	100.0	0

Relationship status was also found to be a factor that may affect the individual's ability to repay their loan, as earlier stated in the literature review. Below is an overview of our sample's relationship status.

Relationship status	Frequency	Percent	Valid percent	Cumulative percent
Single	4995	97.9	97.9	97.9
Married	70	1.4	1.4	99.3
Single parent	27	.5	.5	99.8
Divorced	5	.1	.1	99.9
Widowed	3	.1	.1	100
Total	5100	100.0	100.0	0

Acquiring a bursary or scholarship was also seen to affect the student's ability and attitude towards repaying back HELB loans. Below is an overview of our sample and the percentages that were awarded these extra means as opposed to those who did not acquire them.

Bursaries	Frequency	Percent	Valid percent	Cumulative percent
Acquired	4646	91.1	91.1	100
Did not acquire	454	8.9	8.9	8.9
Total	5100	100.0	100.0	0

Gender was found to be a huge determinant of whether one defaulted on their loans or not. Men were found to have a higher likelihood to default than women as discussed in the literature review. Below is a table that shows the variation of gender in our sample.

Gender	Frequency	Percent	Valid percent	Cumulative percent
Male	2849	55.9	55.9	55.9
Female	2251	44.1	44.1	100
Total	5100	100.0	100.0	0

3.1.2 Target Population

A target population is the entire group of individuals for which the survey data are to be used to make inferences for the study. It defines those individuals for which the findings of the survey are meant to generalize. The entire population included HELB financial statements and data from 1995, when its operation went nation wide, to 2014. Since the sample population was too large, raw and unpolished, our study only took data from 2009 to 2014 as this is the period when the Board had began experiencing major improvements in their disbursement and recovery policies.

For this study, we focused on individuals who had completed their higher education studies from within the first year of completion upto 50 years since completion. The survey therefore includes individuals from ages 23 to 75 who both had and had not completed paying off their student loans.

3.2 Sources of Data

In this study, we used secondary data obtained from various sources including financial statements and reports at the Kenyan Higher Education Loans Board (HELB). The study sample consists of Kenyan students who studied both in private and public universities and colleges, and had benefited from the loans.

The six year period was chosen because it is more current and it was a time when the Board had made major changes and were experiencing better results from their operations. This period is therefore expected to yield a good representative result. Data analysis was carried out with the aid of both descriptive and inferential analysis.

3.2.1 Data Analysis

The data was analyzed using, firstly, Visual Basic for Applications (VBA) which is found in Microsoft Excel. VBA is a programming language that allows the user to create their own defined and customized functions. This was necessary especially in the initial stage of polishing the large amount of information in the data in order to gather a sample where only the relevant information was present.

The polished data sample was entered into R Studio to build the multiple logistic regression model. This required a number of steps including creating dummy variables for the loan amount and the number of days for which the applicant had delayed their payments for. This was because these variables were categorical and we had to find a way to include them into the model as continuous variables. Hair et al., (2010) found that categorical variables can be represented as dummy variables and included in the analyses requiring only continuous variables. This method was applied in this study, where any categorical variable was made into a dummy variable for ease of functioning of the model.

3.2.2 Variable Selection

In the discussion of regression, one may assume that the explanatory variables in the model are selected in advance. However, the selection of these variables is part of the analysis since they are usually not predetermined. The two main approaches to variable selection are; the regressions approach and by use of automatic methods.

For this analysis, we used the regressions approach because it considers all possible subsets of the pool of explanatory variables and finds the model that best fits the data according

to a given criteria. The criteria used for this study is the Akaike Information Criterion (AIC), which assigns scores to each model and allows us to choose the model with the best score. The lower the AIC compared to the null deviance, the better the model will be.

We used the *step* function to perform variable selection. To do this, we began by specifying a starting model and the range of models which we want to examine in the search. One may use forward selection, backward selection or both, which is the one we went ahead with.

The *step* function comes before defining the logistic regression model in R. For the "both" selection, the function starts with the variable that has the highest information criteria: AIC = 6257.05 and eliminates it, and continues on to the second highest AIC = 6255.05 , and so on until the lowest possible AIC = 6236.99 is present.

The variables present where there was the lowest AIC were our best selection of inferential variables, in this case it was the loan amount and the father being alive. These two variables were chosen by the model to have the lowest correlation since they had the highest effect to the model. We found that both loan amount and the father of the beneficiary being alive influenced the account status positively. We also saw that the coefficient of loan amount was significant, ($p < 0.05$), while that of the father being alive was not as significant.

We used the *newdat* function in R to show the prediction of a student's account status given their loan amount. The loan amount here is rounded to the nearest thousand for ease of plotting. The graph below shows the relation between the loan amount and the student's account status. We see that as the loan amount increases, so does the likelihood of default

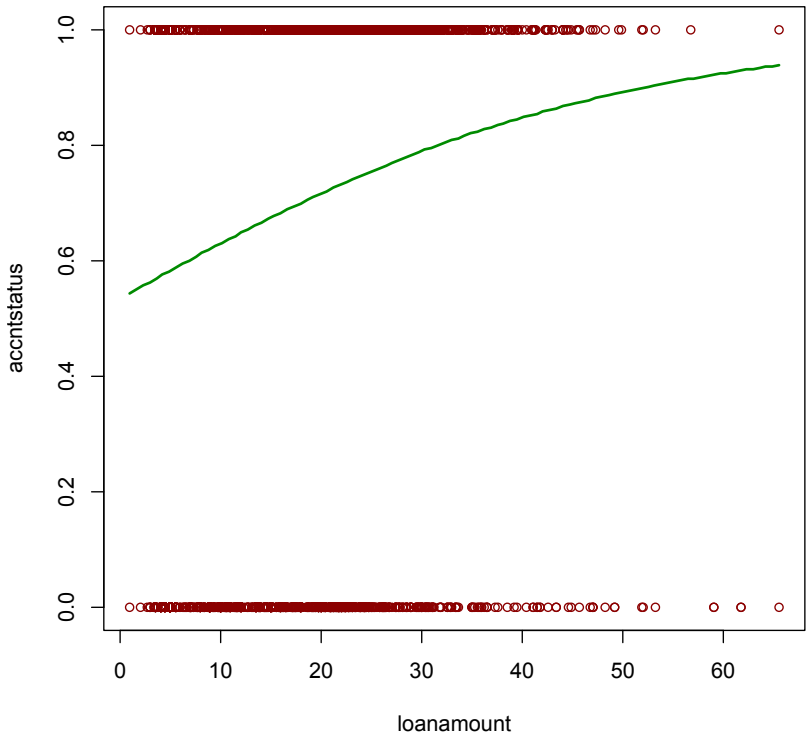

Figure 1: Graph of account status against Loan amount.

3.3 The Model

The dependent variable in logistic regression is dichotomous, meaning it can take the value 1 or 0 with a probability of defaulting and repaying respectively. This type of variable is called a binary variable. As mentioned earlier, predictor variables can take any form i.e. multiple logistic regression does not make any assumptions on them. They need not be normally distributed, linearly related or of equal variance within each category.

Taking our binary outcome as Y with covariates $X_1, ...X_p$, the logistic regression model assumes that;

$$\log\left(P(Y=1 \mid X_1, ...X_p)\right) = \log\left(\frac{P(Y=1 \mid X_1, ...X_p)}{1-P(Y=1 \mid X_1, ...X_p)}\right) = \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p \tag{1}$$

In terms of probabilities this is written as;

$$P(Y=1 \mid X_1, ...X_p) = \frac{\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p\right)}{1+\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p\right)} \tag{2}$$

The unknown model parameters β_o through to β_p are the coefficients of the predictor variables estimated by maximum likelihood, and X_1 through to X_p are the distinct independent variables.

The right hand side of equation (1) above looks similar to a multiple linear regression equation. However, the method used to estimate the regression coefficients in a logistic regression is different from the one use to estimate regression coefficients in a linear regression model.

In logistic regression, coefficients derived from the model, for example β_1 indicate the change in the expected log odds relative to one unit change in X_1, holding other predictors constant. This means that the antilog of an estimated regression coefficient gives us an odds ratio.

Given that unemployment is the greatest cause of student loan default on a global scale, we chose not to give it too much focus in this particular study so that it may not give us extreme or exaggerated values. To establish the default of higher education loans, we have a regression analysis considering the variables; loan amount, employment, age, gender, both parents being alive and employed, whether the beneficiary had acquired a bursary or scholarship, number of dependents and , number of overdue days. The model will be given by the equation below;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{3}$$

where;

$$\beta_0 = Intercept \tag{4}$$

$$\beta_p = coefficients \tag{5}$$

$$X_p = Predictors \tag{6}$$

$$\epsilon = Errorterm \tag{7}$$

We also checked the strength of the model by conducting an Analysis of Variance test. The significance value on the Analysis of Deviance table was tested at 95 percent confidence level and 5 significant levels. The test showed that the model is very strong.

3.3.1 Odds and Log of Odds

Odds express the likelihood of an event occurring relative to the likelihood of it not occurring. Say p is the probability of the event of default occurring, and is given by p = 0.44, then the probability of repaying is 1-0.44 = 0.56. The odds of defaulting will be given by;

$$odds = \frac{P}{1-P} = \frac{0.44}{1-0.44} = 0.79 \tag{8}$$

This implies that the odds of defaulting is 0.79 to 1, and the odds of repaying is 1.27 to 1. Logistic regression uses the log of the odds ratio rather than the odds ratio itself, therefore;

$$Log\ of\ odds = \log\left(\frac{0.44}{1-0.44}\right) = \log\left(\frac{0.44}{0.56}\right) = -0.1047 \tag{9}$$

and so on for other probabilities.

We carried out a crude and an adjusted odds ratio in R. The adjusted odds ratio is the crude odds ratio modified or adjusted to take into account data in the model that could be important. The table below shows the results we got.

	Crude odds	Adjusted odds
Variable in percentages	OR, 2.5 to 97.5	OR, 2.5 to 97.5
loan amount	1.60, 0.02 to 113.94	1.60, 0.02 to 113.76
Father alive	1.12, 0.41 to 3.09	1.12, 0.41 to 3.09

3.3.2 Deviance

Deviance is specifically useful for model selection. We see two types of deviance in our outcome, namely null and residual deviance. The residual deviance is a measure of lack

of fit of the model taken a a whole while the null deviance shows how well the dependent variable is predicted by a model that includes only the intercept.

In our results, we have a null deviance of 6360.5 on 5099 degrees of freedom. The independent variables being included resulted in the decrease of the residual deviance to 6227.1 on 5088 degrees of freedom. The residual deviance reduced by 133.4 with a loss of 11 degrees of freedom.

3.3.3 Fisher Scoring

Fisher scoring iteration is concerned with how the model was estimated. An iterative approach known as Newton-Raphson algorithm is used by default in R for logistic regression. The model is fit based on an approximation about what the estimates might be. The algorithm searches to find out if the fit can be improved by using different estimates instead. If so, it engages in that direction using higher values for the estimates and fits the model again. The algorithm quits when it perceives that searching again would not yield any additional improvement. In our model, we had 4 iterations before the process quit and output the results.

3.3.4 Hosmer-Lemeshow Test

The strength of the model was tested by use of the Hosmer-Lemeshow goodness of fit test. This test evaluates the goodness of fit by initializing several ordered groups of variables and then comparing the number in each observed group to the number predicted by the logistic regression model. Therefore, the test statistic is a chi-square statistic with a desirable outcome of non-significance, meaning that the model predicted does not differ from the one observed.

The ordered groups are created according to their estimated probability where those with the lowest probability are placed in one group and those with higher probability in different groups, up to the highest one read. These groups are further divided into two groups based on the actual observed outcome variable i.e. defaulter or re-payer. The

expected frequencies are obtained from the model.

If the model is strong, then most of the variables with success are classified in the higher deciles of risk and those with failure in the lower deciles of risk.

The Hosmer-Lemeshow goodness of fit test gave us df = 8 and a p-value of less than 2.2e-16, which is very small and definitely less than 0.05, meaning that our model fit the data.

3.4 Model Assumptions

A number of assumptions are made for the multiple logistic regression to function including, there should be no outliers, high leverage values or highly influential points. This assumption is likely not to be followed since we are dealing with data that is stochastic and not normally distributed. Multiple logistic regression also assumes non-perfect separation. If the groups of the outcome variable are perfectly separated by the predictor(s), then unrealistic coefficients will be estimated and effect sizes will be greatly exaggerated.

Furthermore, there needs to be an independence among the dependent variable choices meaning that the choice of or membership in one category is not related to the choice or membership of another category (i.e. dependent variable). This assumption of independence can be tested with the Hausman-McFadden test (Starkweather and Kay, 2012).

3.4.1 Multicollinearity

Multicollinearity occurs when you have two or more independent variables that are highly correlated. This results in problems with understanding which variables contribute to the explanation of the dependent variable, which leads to complications in calculating a multiple logistic regression. It reduces the model's legitimacy and predictive power. To ensure the model is well specified and functioning properly, there are tests that can be run. Variance Inflation factor is one such tool used to reduce multicollinearity.

3.4.2 Variance Inflation Factor (VIF)

This helps to identify the severity of any multicollinearity issues in order for the model to be adjusted accordingly. It measures how much the variance of an independent variable is affected by its interaction with other independent variables. VIFs are usually calculated by the software as part of the regression analysis.

VIFs are calculated by taking a predictor variable, X_i and regressing it against every other predictor variables in the model. This gets you the unadjusted R-squared values which can then be injected into the VIF formula. In the formula below, "i" is the predictor you are looking for;

$$VIF = \frac{1}{1 - R_i^2} \tag{10}$$

The variance inflation factor ranges from 1 upwards, where the numerical value, in decimal form, informs us the percentage the variance is inflated for each coefficient. For instance, a VIF of 1.065709 tells us that the variance of a particular coefficient is 6.5709 percent larger than what we would expect if there was no correlation with other predictors.

Generally, a VIF of 1 indicates zero correlation, if the VIF is between 1 and 5 then there is moderate correlation and anything greater than 5 indicates a high level of correlation. In our sample data, the VIF is as follows; loan amount = 1.001370, employment = 1.008483, age = 1.001269, gender = 1.026480, father alive = 2.981585 , mother employed = 1.064755, mother alive = 3.011166, bursary = 1.065709, dependents = 1.009704, overdue days = 1.152670.

The variance between the coefficients used to build the model were only moderately correlated, therefore our model is without extreme multicollinearity.

3.4.3 Presence of outliers

Outliers are observations identifiable as distinctly separate from majority of the sample, (Hair et al., 2010). The study developed two box plots of account status against the loan amount given to the student, and as well against the number of overdue days that the

individual had delayed their payments.

The outliers on both of them were quite extreme, especially small amounts ranging from 700 to 4,200 shillings on the one showing loan amounts. This indicates that the individuals had very little loan left to clear but had not yet done so and this amounts remained dormant on their accounts, and are now revealed as outlier variables. The whiskers on the box plots were longer than the size of the box itself. A well proportioned tail would produce whiskers about the same length as the box, or slightly longer.

The box plot for defaulters is slightly bigger than that of non-defaulters indicating the difference between the highest loan amount to the lowest is larger for the defaulters than it is for their counterparts. The median on the defaulter's box plot is visually equidistant from the upper quatile to the lower quatile, meaning that loan defaulters are well spread whether they took a larger loan amount or a smaller loan amount. However, for the non-defaulters, the number of individuals who took up larger loans are closer together than those who took lower amounts in loans.

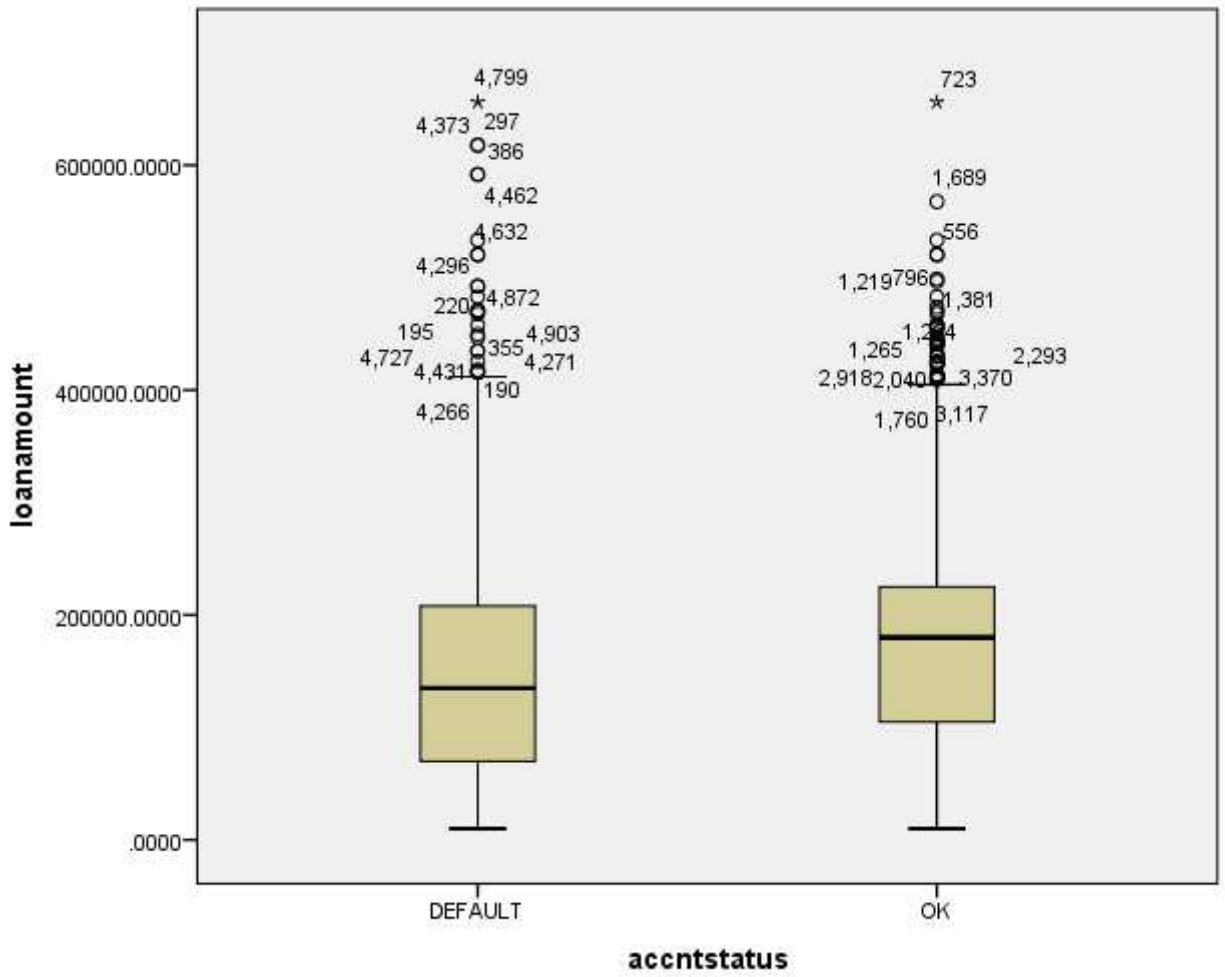

Figure 2: Box plot of Loan amount against account status.

The box plot on overdue days showed that the majority of beneficiaries delayed their payments by about 50 days. For the non-defaulters, the box plot is very short meaning that there is certain agreement with taking a shorter number of days to pay off the loans as opposed to taking long. This is contrary to the defaulters box plot which is longer and more evenly spread.

The outliers on these two box plots tell the tale of those individuals who completed school a very long time ago and have not yet cleared their student loans. They are the extreme values indicated above the whiskers.

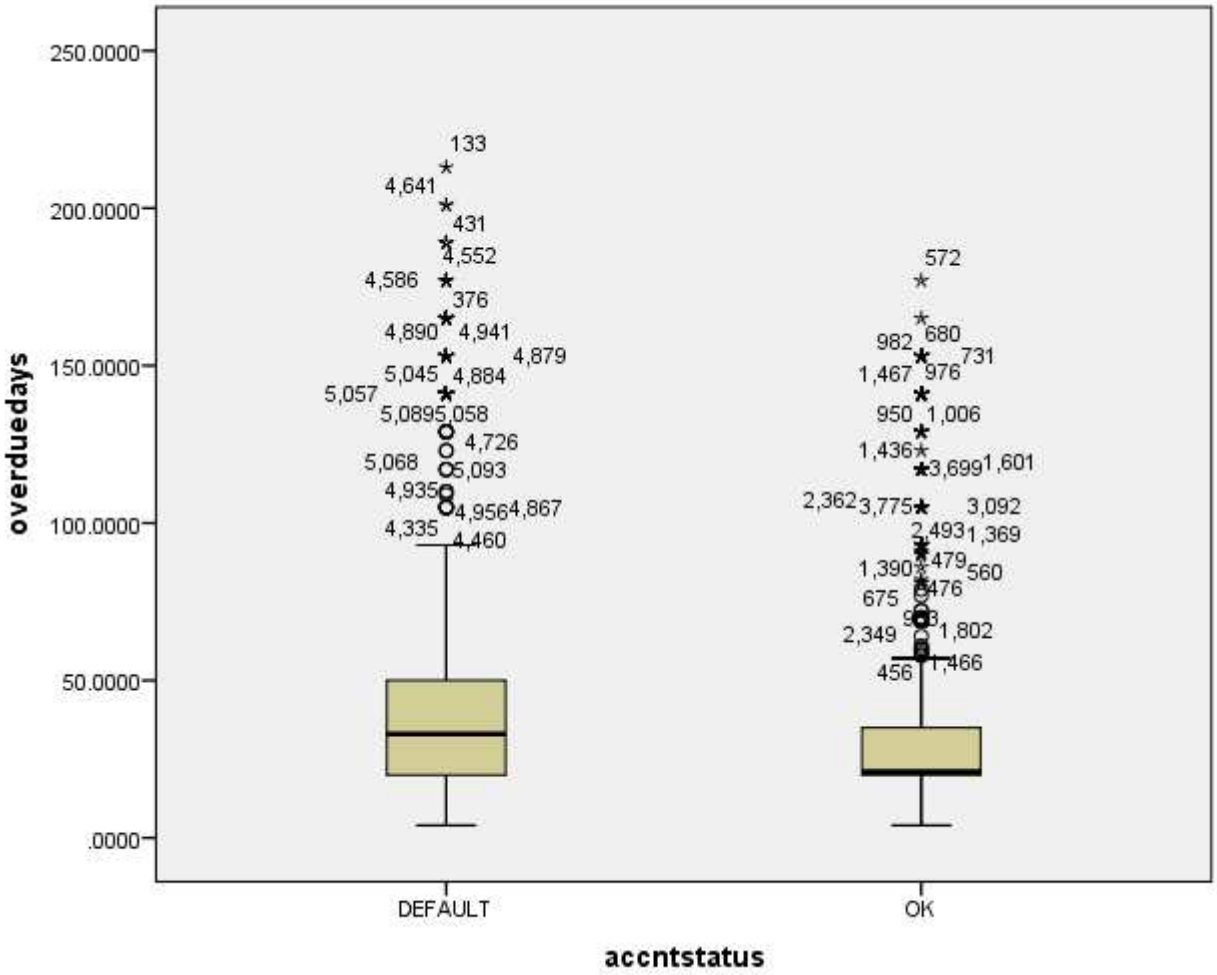

Figure 3: Box plot of Overdue days against account status.

To treat the outliers situation, we converted the variables in the sample population into probabilities. This allowed for ease of estimation and guaranteed lower errors in the model fit. Converting the variables into probabilities also allowed us to properly gauge the likelihood that an individual had certain characteristics that led them to default.

Below is a bar chart of frequency against age. The chart shows the point in an individual's life when he or she is most likely to default. The chart also shows the frequency of people at that age who are most likely to default. The most frequent ages lie between 23 and 40 years of age. This is because at this age, most people have completed their studies and have ventured into the work force. At this age is when most people have many responsibilities, including career and family obligations. This may contribute to

their default on student loans.

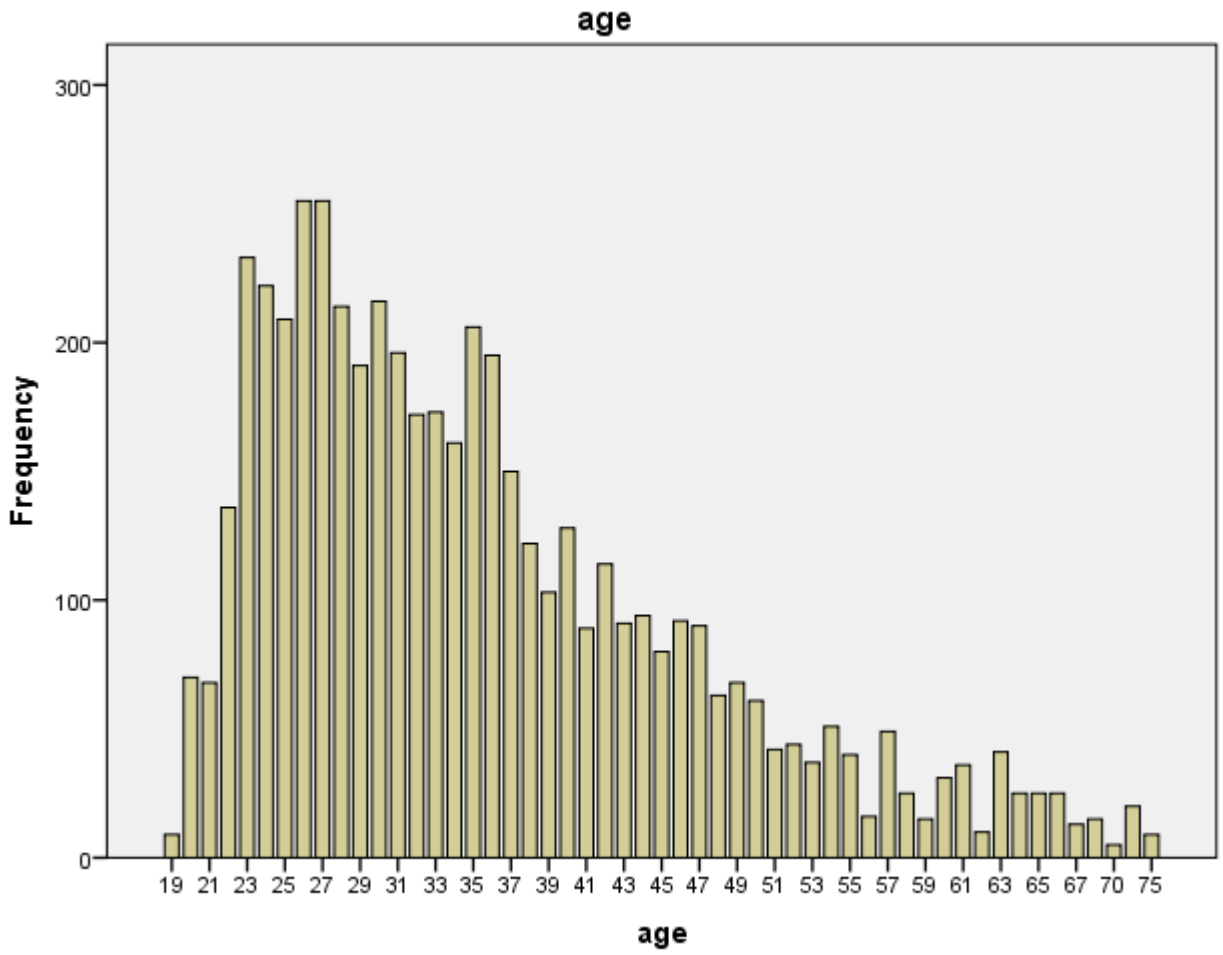

Figure 4: Bar Chart of Frequency against age.

4 Research Findings

One of the main objectives of this research was to develop a quantitative model that returns an individual's risk of default. This model can be used by HELB to categorize new loan applicants as highly likely to default or not likely to default. Multiple logistic regression was developed using the standardized coefficients which are the multiplier of the independent variables and their predictors. Based on the summary of the logistic

regression presented in the table below, the most significant variable in the model was the loan amount. Using the predictors and their coefficients, the logistic regression equation is given as below;

$$Y = 0.0899 + 0.03959\text{loan amount} + 0.13174\text{employment} - 0.13433\text{age} - 0.17722\text{gender} + 0.37899\text{father alive} - 0.07674\text{mother employed} - 0.06822\text{mother alive} - 0.10349\text{bursary} + 0.00432\text{dependents} - 0.0732\text{overdue days}$$

The coefficients above indicate the partial contribution of each variable to the regression equation by holding other variables constant.

Coefficients	Estimate	Std. Error	z-value	Pr($>\|z\|$)
(Intercept)	0.089900	0.364316	0.247	0.805
loanamount	0.039585	0.003625	10.921	$< 2e-16$
employment	0.131739	0.167422	0.787	0.431
age	-0.134330	2.201376	-0.061	0.951
gender	-0.177216	0.531190	-0.334	0.739
f.alive	0.378986	0.314789	1.204	0.229
m.employed	-0.076737	0.154675	-0.496	0.620
m.alive	-0.068219	0.367307	-0.186	0.853
bursary	-0.103491	0.135569	-0.763	0.445
dependents	0.004320	0.569261	0.008	0.994
overduedays	-0.073199	0.066701	-1.097	0.272

5 Discussion and Conclusions

This study went into finding out what causes students of higher education to default on their loans. Personal characteristics and attributes were found to be key variants, with unemployment being the highest by far. Since it is apparent to say that unemployment or lack of lucrative employment is the major cause of student loan default, we placed more focus on the other variants.

The findings of the study with regards to cumulative amount of loan given to the student and default indicated a positive relationship indicated by the significance of its p-value. We saw that students who took up loans more frequently ended up with a huge loan at the end of their studies, which they had to pay back but with little or no means to do so especially given the unemployment rates in the country. This was in line with the study done by (Choy and Li,2006; Dynarski, 1994 and Lochner and Monge-Naranjo, 2004), who found that the larger the loan the higher the likelihood of default.

The findings indicated that if HELB monitored how much money cumulatively they reimbursed to applicants, they would be able to categorize separately those who would default from those who would be less likely to default. Typically, the greater the debt accumulated over time, the more likely one is to default.

The average loan amount advanced to defaulters was KES 93,432.13 with a maximum and minimum of KES 240,000 and 20,000 respectively. The standard deviations of the loan amounts and the study period are indicative that for each additional half year, loan amounts of KES 47,990.20, on average, had been disbursed to individual defaulters in the course of their study periods between 2009 and 2012, (Lidoroh, Determinants of Student Loan Default in Kenya, 2012).

The number of overdue days played a huge role in contributing to their likelihood to default where 73 percent of individuals with over 150 days overdue were highly likely to default than individuals with less than that. This is because their loan continues to accumulate interest as the days add up, which is one of HELB's initiatives for loan recovery

i.e. charging a penalty to those individuals who are late on their payments. This could make a defaulter out of an individual who would otherwise not fall into default, especially due to the fact that the employment is always fluctuating with the economy.

Students who had both parents, even if the parents were not both employed, showed a significant ability to not default on their loans by 68 percent compared to orphaned loan beneficiaries. Additionally, these individuals showed greater persistence in servicing their loans in due time.

Given the Logistic Regression formula for probability of success or failure, we should be able to find the probability of default, P, by keying in details into the model equation. The details are the $\beta's$ which we found through model simulation in R Studio.

As expected, individuals with variable probabilities that favor default tendencies will be more likely to default. For instance an individual who had more overdue days, is older, orphaned and took a huge amount of loan is more likely to default than a counterpart with opposite qualities to these. We can find this out by keying in each individual's unique probabilities to the model equation to find out their particular probability. For example,

6 Limitation and Recommendations

The major limitation of this study is the lack of exhaustive data variables of interest i.e. time to defaulting. Even though we are immensely grateful to HELB for the data provided to us, the best kind would have been one that shows the time until the first time a student defaults, as well as how many times a student's default tendencies recur. This would have been perfect for the analysis of all the exact events that lead to the first time defaulting.

Future potential research area involves modeling time to default for both single event and recurrent events. This will enable computation of hazard functions and rates. Another potential area of study is on how to treat outliers in this setting.

Manuscript

A manuscript entitled **Modeling factors affecting Probability of Loan Default: A Quantitative Analysis of the Kenyan Students' Loan** authored by Pauline N. Kamau, Lucy Muthoni and Collins Odhiambo will be submitted for editorial review at **Science Journal of Applied Mathematics and Statistics** before 31st May, 2018.

References

[1] Nick Hillman, Don Hossler, Jacob P.K. Gross & Osman Cekic *What Matters in Student Loan Default: A Review of the Research Literature* Journal of Student Financial Aid, Issue 1, Article 2, 1-10-2010

[2] Blom, Andreas, Reehana Raza, Crispus Kiamba, Himdat Bayusuf, and Mariam Adil. 2016. *Expanding Tertiary Education for Well-Paid Jobs: Competitiveness and Shared Prosperity in Kenya.* World Bank Studies. Washington, DC: World Bank. doi:10.1596/978-1-4648-0848-7. License: Creative Commons Attribution CC BY 3.0 IGO

[3] Anamaria Felicia Ionescu *The Federal Student Loan Program: Quantitative Implications for College Enrollment and Default Rates* Economics Faculty Working Papers, Colgate University Libraries, Summer 6-2008

[4] Felicia Ionescu & Nicole Simpson *Default Risk and Private Student Loans: Implications for Higher Education Policies* Finance and Economics Discussion Series, 2014-066

[5] Maja Pohar, Mateja Blas & Sandra Turk *Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study*

[6] Michal T. Njenga *The Determinant of Sustainability of Student Loan Schemes: Case Study of Higher Education Loans Board*
Scool of Business, University of Nairobi, November 2014

[7] Mwangi Johnson Muthii *Predicting Student's Loan Default In Kenya: Fisher's Discriminant Analysis Approach* School of Mathematics, University of Nairobi, 2015

[8] Emile A.L.J. van Elen *Term structure forecasting*
School of Economics and Management, Tilburg University, 2010

[9] Peter C., B. Phillips & Jun Yu *Maximum Likelihood and Gaussian Estimation of Continuous Time Models in Finance* Cowles Foundation for Research in Economics, Yale University, University of Auckland and University of York.

School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903

[10] Stephen Crowley *Maximum Likelihood Estimation of the Negative Binomial Distribution* Unpublished Working Paper, 2012

[11] Elizabeth Herr & Larry Burt *Predicting Student Loan Default for the University of Texas at Austin*

[12] Christophe Hurlin *Maximum Likelihood Estimation and Geometric Distribution* Advanced Econometrics, University of Orleans, 2013

[13] Mark Huggett, Gustavo Ventura & Amir Yaron *Sources of Lifetime Inequality* American Economic Review 101, 2923-2954, 2011

[14] Newey, Whitney K. & Daniel McFadden. *Large sample estimation and hypothesis testing*. Handbook of econometrics, Vol. 4, 1994

[15] Stu Field *Parameter Estimation via Maximum Likelihood*. Unpublished working paper, 2009

[16] Konstantin Kashin *Statistical Inference: Maximum Likelihood Estimation*. Journal of Finance, Spring 2014